Statistics - Measures of central tendency and dispersion Class 2
Session 2

Oscar BARRERA
oscardavid.barrerarodriguez@sciencespo.fr

January 31, 2018

Before starting

Some important tips of notation

	Measure of...	Population Parameter	Sample Statistic	Parameter Estimate
Mean	Arithmetic average	μ	$\overline{X}$ or M	$\overline{X}$ or M
Population Size	Number of subjects	N	n	n
Variance	Average variability	σ^2	s^2	$\hat{s}^2$
Standard Deviation	Average distance between a value and the mean	σ	s	$\hat{s}$
Correlation	Strength of relationship between two variables	ρ	r	r

Measures of Central Tendency

So you collected data, created a frequency distribution, made a graph...now what?

- It would be nice to have a single value to represent a distribution a **"descriptive statistic"**

Descriptive statistic

- Is a value that describes or represents a set of data.
  - But what value from a distribution should be used as the representative?
  - Why is any one value in the data set better than another value?

There are at least **three characteristics** you look for in a descriptive statistic to represent a set of data.

Measures of Central Tendency

There are at least **three characteristics** you look for in a descriptive statistic to represent a set of data.

1. Represented: A good descriptive statistic should be similar to many scores in a distribution. (High frequency)
2. Well balanced: neither greater-than or less-than scores are overrepresented
3. Inclusive: Should take individual values from the distribution into account so no value is left out

Measures of Central Tendency

In a normal distribution the most frequent scores cluster near the center and less frequent scores fall into the tails.

Central tendency means most scores(68%) in a normally distributed set of data tend to cluster in the central tendency area. [Come back to characteristics!]

The Mode

The mode (Mo) is the most frequently occurring score in a distribution.
Example with a set of quiz scores (X):

10 10 10 9 9 9 9 8 8 8 8 8 7 7 7 7 7 7 6 6 6 6 6 6 6 5 5 5 5 5 5 5 5 5 4 4 4 4 4 4 4 4 3 3 3 3 3 2 2 1

The Mode

The mode (Mo) is the most frequently occurring score in a distribution.
Example with a set of quiz scores (X):

10 10 10 9 9 9 9 8 8 8 8 8 7 7 7 7 7 7 6 6 6 6 6 6 6 5 5 5 5 5 5 5 5 5 4 4 4 4 4 4 4 4 3 3 3 3 3 2 2 1

What is the Mo?

- X = 5 has the greatest frequency (9); hence, the mode is 5.

The Mode
Limitations

- **Multi-modal:** There can be more than one mode
- **Lack of Representativeness** :It may not be a good representative of all values

The Mode

Limitations

- **Multi-modal:** There can be more than one mode

Examples

10 10 9 9 9 9 9 9 8 8 7 7 6 6 5 5 4 4 4 4 4 4 3 3 2 2 1 1

Two most frequent scores, X = 9 and X = 4,

- Or in case of a rectangular distribution

10 9 8 7 6 5 4 3 2 1 0

The Mode

Limitations

- **Lack of Representativeness** :It may not be a good representative of all values

Examples

10 10 10 10 10 10 10 9 8 7 6 5 4 3 2 1 0

The mode might be at one end of a distribution, not at the center

The Mode

Limitations

It is possible that the most frequent score has a frequency that is only one or two counts greater than the second most frequent score. In the first example above the mode was X = 5. However, X = 4 had a frequency of eight, which is only one less than the frequency of nine for X= 5.
Why is 5 a better mode than 4? There is no answer.

The Median

The median (Md) is the middle score of a distribution.

- Half on the left half on the right (the 50th percentile.)
  - Better measure of central tendency than the mode since it balances perfectly distribution.

How to find it?

The Median
How to find it?

Two simple steps

1. determine the median's location
2. find the value at that location.
    - It differs whether you have an even or an odd number of scores.

With ODD number of observations. For example: n=11

10 10 9 7 7 6 5 4 3 2 2

Use the following equations

$$Md = \frac{N+1}{2}$$

Solving with N=11?

The Median

How to find it?

With ODD number of observations. For example: n=11

10 10 9 7 7 6 5 4 3 2 2

Use the following equations

$$Md = \frac{N+1}{2}$$

Solving with N=11?

$$Md = \frac{11+1}{2} = 6th$$

This indicates the median is located at the sixth position in the distribution.

The Median

How to find it?

$$Md = \frac{11+1}{2} = 6th$$

This indicates the median is located at the sixth position in the distribution.

- Order the data: rank-ordered from smallest to largest.
- Count off six positions starting with the smallest value.

X:	*2*	*2*	*3*	*4*	*5*	*6*	*7*	*7*	*9*	*10*	*10*
Ordinal Position:	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th

The Median

How to find it?

Two simple steps

1. determine the median's location
2. find the value at that location.
   - It differs whether you have an **even** or an **odd** number of scores.

With EVEN number of observations. For example: n=12

20 19 18 16 15 14 12 11 11 11 10 9

You have to determine the two positions around the median

$$Md = \frac{N}{2}, \frac{N+2}{2}$$

Solving with N=11?

The Median

How to find it?

With EVEN number of observations. For example: n=12

20 19 18 16 15 14 12 11 11 11 10 9 Use the

You have to determine the two positions around the median

$$Md = \frac{N}{2}, \frac{N+2}{2}$$

Solving with N=12?

$$Md = \frac{12}{2}, \frac{12+2}{2} = Md = 6th, 7th$$

The Median
How to find it?

With EVEN number of observations. For example: n=12

20 19 18 16 15 14 12 11 11 11 10 9 Use the

You have to determine the two positions around the median

$$Md = \frac{12}{2}, \frac{12+2}{2} = Md = 6th, 7th$$

Thus, the median is between sixth and seventh positions.

X:	9	10	11	11	11	12	14	15	16	18	19	20
Ordinal Position:	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	20th

The Median

How to find it?

With EVEN number of observations. For example: n=12
You have to determine the two positions around the median

$$Md = \frac{12}{2}, \frac{12+2}{2} = Md = 6th, 7th$$

Thus, the median is between sixth and seventh positions.

X:	9	10	11	11	11	12	14	15	16	18	19	20
Ordinal Position:	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	20th

The average of 12 and 14 is $(14 + 12)/ = 13$, which is the median value.

The Median

Limitations

Some limitations of the Median:

- The median does not take into account the actual values of the scores in a set of data.

Example:take the following two sets of scores, each with n = 5 scores:

Set 1:	1	4	5	7	10
Set 2:	4	4	5	6	6

Median is equal to 5: which is unaffected by the values greater and less than its value.

The Median

Limitations

Some limitations of the Median:

- The median does not take into account the actual values of the scores in a set of data.

Median is unaffected by the values greater and less than its value.

However: it is frequently reported as a measure of central tendency when data sets are incomplete or data are severely skewed.

Example?

Quartiles

Quartiles split a distribution into fourths or quarters of the distribution.

- There are actually three quartile points in a distribution:
  - The 1st quartile (Q1) separates the lower 25% of the scores from the upper 75% of the scores;
  - the 2nd quartile (Q2) is the median and separates the lower 50% of the scores from the upper 50%;
  - and the 3rd quartile (Q3) separates the upper 25% of the scores from the lower 75%.

Steps: Determining the quartile value is like determining the median.

Quartiles
How to find it

When determining the median, there are different procedures for determining the quartiles based on your having an **even** versus an **odd** number of scores.

- For the 1st quartile ODD $Q_1 = \frac{N+1}{4}$
- For the 1st quartile EVEN $Q_1 = \frac{N+2}{4}$
- For the 3st quartile ODD $Q_3 = \frac{3N+3}{4}$
- For the 3st quartile EVEN $Q_3 = \frac{3N+2}{4}$

Quartiles

How to find it

Example: ODD number. n=9

7 7 8 9 10 11 11 14 15

Then The locations of the first and third quartiles are:

$$Q_1 = \frac{9+1}{4} = 2,5thpos$$

$$Q_3 = \frac{3*9+3}{4} = 7.5thpos$$

Quartiles

How to find it

Then The locations of the first and third quartiles are:

$$Q_1 = \frac{9+1}{4} = 2,5thpos$$

$$Q_3 = \frac{3*9+3}{4} = 7.5thpos$$

Position:	1st	2nd	3rd	4th	5th	6th	7th	8th	9th
Score:	7	7	8	9	10	11	11	14	15

How to estimate Q_1 and Q_3 ?

Quartiles

How to find it

Then The locations of the first quartiles is:

$$Q_1 = \frac{9+1}{4} = 2,5thpos$$

Position:	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}	9^{th}
Score:	7	7	8	9	10	11	11	14	15

$Q_1 = 2.5$ means the first quartile is the average of the values at the second and third positions. $(7+8)/2 = 7.5$

Quartiles

How to find it

Then The locations of the third quartiles is:

$$Q_3 = \frac{3 * 9 + 3}{4} = 7.5thpos$$

Position:	1st	2nd	3rd	4th	5th	6th	7th	8th	9th
Score:	7	7	8	9	10	11	11	14	15

$Q_3 = 7.5$ means the third quartile is the average of the values at the seventh and eight ordinal positions $(11 + 14)/2 = 12.5$.

Quartiles
How to find it

Example: EVEN number. n=10

6 7 7 8 9 10 10 11 11 14

Then The locations of the first and third quartiles are:

Quartiles

How to find it

Then The locations of the first and third quartiles are:

$$Q_1 = \frac{10+2}{4} = 3thpos$$

$$Q_3 = \frac{3*10+2}{4} = 8thpos$$

Position:	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Score:	6	7	7	8	9	10	10	11	11	14

Hence $Q1 = 7$. and $Q3 = 11$. The interquartile range is $Q3 - Q1 = 11 - 7 = 4$.

The Mean

Understanding Sigma

The symbol Σ in Maths means **Summation**. Means to add all values to the right of Σ (say var X).

$$\Sigma X$$

Means to sum up all values that belong to variable X.
Example!!

Professor	2005 Salary (Y)	2006 Salary (X)
Dr. Java	39000	41000
Dr. Spock	43500	46000
Dr. Evil	48000	51000
Dr. Griffin	52500	56000
Dr. Who	57000	62000

Estimate ΣX, ΣY and $\Sigma DrEvil$

The Mean

Understanding sigma

Professor	2005 Salary (Y)	2006 Salary (X)
Dr. Java	39000	41000
Dr. Spock	43500	46000
Dr. Evil	48000	51000
Dr. Griffin	52500	56000
Dr. Who	57000	62000

Estimate ΣX , ΣY and $\Sigma Dr.Evil$

$$\Sigma X = 39,000 + 43,500 + 48,000 + 52,500 + 57,000 = 240000$$

$$\Sigma Y = 41,000 + 46,000 + 51,000 + 56,000 + 62,000 = 256000$$

Understanding sigma

IMPORTANT!!

Σ is a grouping symbol like a set of parentheses and everything to the right of Σ must be completed before summing the resulting values.

Example

X	Y
10	5
9	5
8	4
6	4
4	3
3	2

Find: ΣX^2, $(\Sigma X)^2$, $\Sigma(X - Y)$, $\Sigma(X - Y)^2$ and ΣXY

Understanding Sigma

$$\Sigma X^2 = 100 + 81 + 64 + 36 + 16 + 9 = 306$$

$$(\Sigma X)^2 = (10 + 9 + 8 + 6 + 4 + 3)^2 = (40)^2 = 1600$$

$$\Sigma(X - Y) = (10-5)+(9-5)+(8-4)+(6-4)+(4-3)+(3-2) = 17$$

$$\Sigma(X - Y)^2 = (10-5)^2+(9-5)^2+(8-4)^2+(6-4)^2+(4-3)^2+(3-2)^2 = 63$$

$$\Sigma XY = 50 + 45 + 32 + 24 + 12 + 6 = 169$$

The mean

Is the arithmetic average of all the scores in a distribution.

- The mean is the most-often used measure of central tendency
  1. It evenly balances a distribution so both the large and small values are equally represented
  2. Takes into account all individual values.

To estimate it.. in two steps

1. add together all the scores in a distribution ΣX.
2. divide that sum by the total number of scores in the distribution.

Sample

$$\overline{X} = M = \frac{\Sigma X}{n}$$

Population

$$\mu = \frac{\Sigma X}{N}$$

The Mean
Example

Calculate the mean from the sample of $n = 11$:

$$2\ 2\ 3\ 4\ 5\ 6\ 7\ 7\ 9\ 10\ 10$$

The Mean
Example

Calculate the mean from the sample of n = 11:

2 2 3 4 5 6 7 7 9 10 10

Solution

$$\overline{X} = M = \frac{\Sigma X}{n} = \frac{65}{11} = 5,909$$

The median = 6. The difference between the mean and the median reflects the mean's taking into account the individual values.

Mean vs Median

Just to emphasize the fact that the mean takes into account all values in a set of data.. Recall this example

Set 1:	1	4	5	7	10
Set 2:	4	4	5	6	6

The mean for each values would be:
Set 1: $\overline{X} = \frac{27}{5} = 5,4$; Set 2 $\overline{X} = \frac{25}{5} = 5$
You can see why the mean is a more accurate measure of central tendency as it takes the individual scores into account; the median does not.

Mean

IMPORTANT:
The mean is defined as the mathematical center of a distribution.(i.e., differences between scores and the mean). The sum of such different IS zero for **any** distribution. **Perfect balancing point**

X	4	4	5	6	6	
$\bar{X}$	5	5	5	5	5	
$(X - \bar{X})$	-1	-1	0	1	1	$\sum (X - \bar{X}) = 0$

Thus the sample mean is the preferred measure of central tendency, because it is the best unbiased estimator of the population mean (μ). If sample was randomly selected then it is representative of the population.

The Mean
Limitations

One problem...it takes individual values into account,

- By taking individual values into account the mean can be influenced by extremely large or extremely small values (outliers).
- Specifically, extremely small values pull the mean down and extremely large values pull the mean up.
- This only occurs in skewed distributions: it is good to use the median as a measure of central tendency (Why?)

What to use, when?

- Nominal data $\Rightarrow$ the mode is the best measure of central tendency. (i.e Sex)
- ordinal scale $\Rightarrow$ the median may be more appropriate. e the individual values on an ordinal scale are meaningless
- interval or ratio scale, $\Rightarrow$ the mean is generally preferred. (Counts for all obs)

The Mean Median and Mode in Normal and Skewed Distributions

The positions of the mean, median, and mode are affected by whether a distribution is normally distributed or skewed.

- In data normally distributed: mean =median = mode.
  - 50% of the scores must lie above the center point, which is also the mode, and the other 50% of the scores must lie below.
  - Because the distribution is perfectly symmetrical the differences between the mean and the values larger than the mean must cancel out.

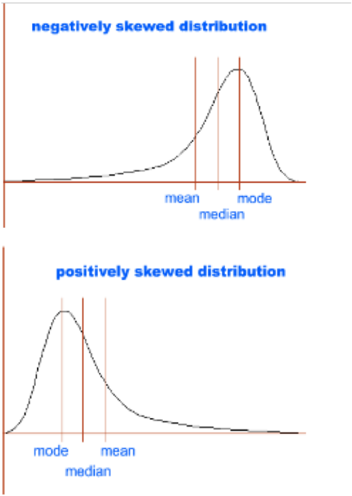
Negatively skewed

- the tail of the distribution is to the left and the hump is to the right.

Positively skewed

- the tail of the distribution is to the right and the hump is to the left,

The Mean Median and Mode in Normal and Skewed Distributions

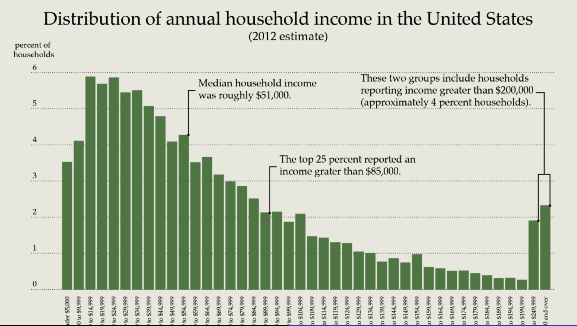


The Mean Median and Mode in Normal and Skewed Distributions

So if the mean differs from the median/mode, the distribution is skewed.

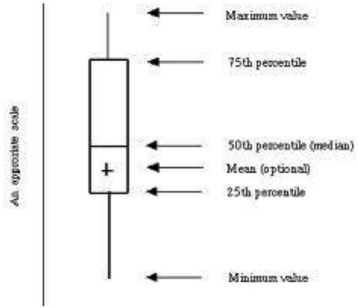
- The median is better as the measure of central tendency when the distribution is positively or negatively skewed.

Best example. Income



Box Plots

is a way to present the dispersion of scores in a distribution by using five pieces of statistical information: the minimum value, the first quartile, the median, the third quartile, and the maximum value.



Box Plots

Examples: two classes who took the same final exam

Statistic	Class A	Class B
Minimum Value	49	50
Q_1	60	58
Median	64	65
Q_3	73	69
Maximum Value	82	75

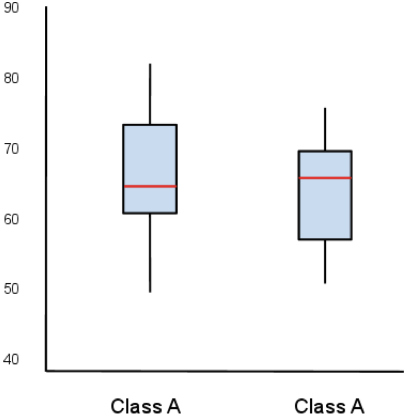
Box Plots

Examples: two classes who took the same final exam

Statistic	Class A	Class B
Minimum Value	49	50
Q_1	60	58
Median	64	65
Q_3	73	69
Maximum Value	82	75

Box Plots

Examples: two classes who took the same final exam



What can it be concluded from the graph?

Box Plots

What can it be concluded from the graph?

- inter quartile range is about the same in each class
- it is shifted down slightly in class B
- The medians are about the same for each class
- overall range of scores is slightly less in Class B then in Class A.

If the median line is equidistant from the edges of the box (i.e., equidistant from Q1 and Q3), then the data is not likely skewed.

- If the median line is closer to the bottom edge of the box (closer to Q1), this suggests positive skew A or B?
- If the median line is closer to the upper edge of the box (closer to Q3), this suggests negative skew, A or B?

What is Variability? Why is Measured?

The variety of colors, shapes, sizes, flavors, and importantly the spiciness of those chili peppers is beautiful!

What is Variability? Why is Measured?

Variability is simply the differences among items, which could be differences in eye color, hair color, height, weight, sex, intelligence, etc.

- If measures of central tendency (mean, median, and mode) estimate where a distribution falls
- Variability measures the dispersion or similarity among scores and tell us about the variety of the scores in a distribution.

Why do we measure variability?

- The mean takes into account all scores in a distribution. Even if you know where the mean falls you do not know anything about the actual scores.
- That is, if I tell you the mean of a set of $n = 10$ scores on a quiz with a range of 0 to 10 is $M = 7$.
- The mean just tells you were a distribution of data tends to fall.

What is Variability? Why is Measured?

- The mean takes into account all scores in a distribution. Even if you know where the mean falls you do not know anything about the actual scores.
- That is, if I tell you the mean of a set of n = 10 scores on a quiz with a range of 0 to 10 is M = 7.
- The mean just tells you were a distribution of data tends to fall.

For example, if the mean for a set of ten scores on a statistics quiz is 7, those ten scores could be:

$$7+7+7+7+7+7+7+7+7+7 = 70/10 = 7$$

or

$$6+6+6+6+6+8+8+8+8+8 = 70/10 = 7$$

or

$$1+1+5+5+9+9+10+10+10+10 = 70/10 = 7$$

measures of variability

There are several measures of variability. I will go from the easiest to more complex ones

- the range
- sum of squares
- the variance
- the standard deviation.

The range

For each of the following sections I will calculate each measure of variability on both of the following sets of n = 10 scores. Note that in both sets the sum of scores and the mean are identical:

Set I:	*2*	*2*	*2*	*4*	*5*	*5*	*6*	*6*	*8*	*10*	*$\sum X = 50$*	$\overline{X} = 5$
Set II:	4	4	5	5	5	5	5	5	6	6	$\sum X = 50$	$\overline{X} = 5$

The range is the largest value minus the smallest value.

- Set I the range is $10 - 2 = 8$
- Set II is $6 - 4 = 2$

It provides information about the area a distribution covers, BUT does not say anything about individual values.

The range

It provides information about the area a distribution covers, BUT does not say anything about individual values.
These two sets have the same rage

- Set A: 1 1 1 1 10
- Set B: 1 2 4 5 10

Clearly there are more differences among scores in Set B than Set A, but the range does not account for this variability.

- What we need is some way to measure the variability among the scores.

Sum of squares

Is the sum of the squared deviation scores from the mean and measures the summed or total variation in a set of data.

$$SS = \Sigma(X - \overline{X})^2$$

- could be equal to zero if there is no variability and all the scores are equal
- Can't be negative (mathematically impossible)

Estimate the SS for the sample:

Set I:	2	2	2	4	5	5	6	6	8	10	$\sum X = 50$	$\overline{X} = 5$
Set II:	4	4	5	5	5	5	5	5	6	6	$\sum X = 50$	$\overline{X} = 5$

Sum of squares

Estimate the SS for the sample:

Set 1			Set 2		
X	$(X - \bar{X})$	$(X - \bar{X})^2$	X	$(X - \bar{X})$	$(X - \bar{X})^2$
10	5	25	6	1	1
8	3	9	6	1	1
6	1	1	5	0	0
6	1	1	5	0	0
5	0	0	5	0	0
5	0	0	5	0	0
4	-1	1	5	0	0
2	-3	9	5	0	0
2	-3	9	4	-1	1
2	-3	9	4	-1	1
		SS = 64			SS = 4

The sum of squares for Set 1 is SS $= 64$ and for Set 2 is SS $= 4$. Larger sum of squares indicates there is more variability among the scores.

Variance

Issues with SS

- Sum of squares measures the total variation among scores in a distribution;
- sum of squares does not measure average variability
- We want a measure of variability that takes into account both the variation of the scores and number of scores in a distribution

Sample Variance S^2

- is the average sum of the squared deviation scores from a mean.
- Measures the average variability among scores in a distribution.

$$S^2 = \frac{\Sigma(X - \overline{X})^2}{n} = \frac{SS}{n}$$

Variance

$$S^2 = \frac{\Sigma(X - \overline{X})^2}{n} = \frac{SS}{n}$$

In our examples?

Set I

$$S^2 = \frac{64}{10} = 6,4$$

Set II

$$S^2 = \frac{4}{10} = 0,4$$

Larger measurements of variance indicate greater variability.

Variance

Issues

The one issue with sample variance is it is in squared units, not the original **unit of measurement**.

- If the original measurements in Sets 1 and 2 were of length in centimeters, when the values were squared the measurement units are now cm2

What is the solution to this issue?

Standard Deviation

Once sample variance has been calculated, calculating the sample standard deviation (s) is as simple as taking the square root of the variance.

$$S = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n}} = \sqrt{\frac{SS}{n}} = \sqrt{S^2}$$

So, in our exercise? Set I

$$S = \sqrt{6,4} = 2,53$$

Set II

$$S = \sqrt{0,4} = 0,63$$

The standard deviation measures the average deviation between a score and the mean of a distribution.
For example, in Set I (Set II), a score is expected to deviate from the mean by 2.530 (0,63).

The Population Formulas for Variability

Calculating sum of squares in a population is no different than calculating the sample sum of squares, the only thing that differs is the symbols.